DESCRIPTION
FULL LENGTH cDNA CLONES AND PROTEINS ENCODED THEREBY

FIELD OF THE INVENTION

The present invention relates to a full length cDNA clone encoding a human protein, a protein encoded by the cDNA clone, and a method for producing them and utilizing them.

BACKGROUND OF THE INVENTION

Currently, the sequencing projects, the determination and analysis of the genomic DNA of various living organisms have been in progress all over the world. The whole genomic sequences of more than 10 species of prokaryotes, a lower eukaryote, yeast, and a multicellular eukaryote, C. elegans are already determined. As to human genome, which is supposed to be composed of three thousand million base pairs, the world wide cooperative projects have been under way to analyze it, and the whole structure is predicted to be determined by the years 2002-2003. The aim of the determination of genomic sequence is to reveal the functions of all genes and their regulation and to understand living organisms as a network of interactions between genes, proteins, cells or individuals through deducing the information in a genome, which is a blueprint of the highly complicated living organisms. To understand living organisms by utilizing the genomic information from various species is not only important as an academic subject, but also socially significant from the viewpoint of industrial application.

However, determination of genomic sequences itself cannot identify the functions of all genes. For example, as for yeast, only the function of approximately half of the 6000 genes, which is predicted based on the genomic sequence, was able to be deduced. As for human, the number of the genes is predicted to be approximately one hundred thousand. Therefore, it is desirable to establish "a high throughput analysis system of the gene functions" which allows us to identify rapidly and efficiently the functions of vast amounts of the genes obtained by the genomic sequencing.

Many genes in the eukaryotic genome are split by introns into multiple exons. Thus, it is difficult to predict correctly the structure of encoded protein solely based on genomic information. In contrast, cDNA, which is produced from mRNA that lacks introns, encodes a protein as a single continuous amino acid sequence and allows us to identify the primary structure of the protein easily. In human cDNA research, to date, more than one million ESTs (Expression Sequence Tags) are publicly available, and the ESTs presumably cover not less than 80% of all human genes.

The information of ESTs is utilized for analyzing the structure of human genome, or for predicting the exon-regions of genomic sequences or their expression profile. However, many human ESTs have been derived from proximal regions to the 3'-end of cDNA, and information around the 5'-end of mRNA is extremely little. Among these human cDNAs, the number of the corresponding mRNAs whose encoding protein sequences are deduced is approximately 7000, and further, the number of full length therein is only 5500. Thus, even including cDNA registered as EST, the percentage of human cDNA obtained so far is estimated to be 10-15% of all the genes.

It is possible to identify the transcription start site of mRNA on the genomic sequence based on the 5'-end sequence of a full length cDNA, and to analyze factors involved in the stability of mRNA that is contained in the cDNA, or in its regulation of expression at the translation stage. Also, since a full length cDNA contains ATG, the translation start site, in the 5'-region, it can be translated into a protein in a correct frame. Therefore, it is possible to produce a large amount of the protein encoded by the cDNA or to analyze biological activity of the expressed protein by utilizing an appropriate expression system. Thus, analysis of a full length cDNA provides valuable information which complements the information from genome sequencing. Also, full length cDNA clones that can be expressed are extremely valuable in empirical analysis of gene function and in industrial application.

A method to synthesize a full length cDNA is known to those skilled in the art. For example, the oligo-capping method (Maruyama K. and Sugano S. (1994) Gene 138: 171-174; Suzuki Y. et al. (1997) Gene 20: 149-156) enables to synthesize a library enriched with full length cDNA, in principle. Once the synthesized cDNA is cloned and the nucleotide sequence is determined, it is possible to estimate whether the cDNA is a full length cDNA clone or not by methods such as the ATGpr (Salamov A.A., Nishikawa T., and Swindells M.B. (1998) Bioinformatics 14: 384-390; http://www.hri.co.jp/atgpr/). However, synthesis efficiency needs to be improved, although it is possible to obtain full length cDNA in a certain probability by combining known methods. It is still difficult to clone a full length cDNA of mRNA that is expressed at very low frequency.

SUMMARY OF THE INVENTION

An objective of the present invention is to provide a novel human protein, a polynucleotide encoding the protein, and their usage.

The inventors have developed a method for efficiently cloning a human full length cDNA that is predicted by the ATGpr etc. to be a full length cDNA clone, from a full length-enriched cDNA library that is synthesized by the oligo-capping method. Then, the inventors determined the nucleotide sequence of the obtained cDNA clones from both 5'- and 3'- ends. By utilizing the sequences, the inventors selected clones that were expected to contain a signal sequence by the PSORT (Nakai K. and Kanehisa M. (1992) Genomics 14: 897-911), and obtained clones that do not contain a cDNA encoding a secretory protein or membrane protein.

The full length cDNA clones of the present invention have high fullness ratio since these were obtained by the combination of (1) construction of a full length-enriched cDNA library that is synthesized by the oligo-capping method, and (2) a system in which the full length ratio is evaluated from the nucleotide sequence of the 5'-end.

Furthermore, the inventors have analyzed the nucleotide sequence of the full length cDNA clones obtained by the method, and deduced the amino acid sequence encoded by the nucleotide sequence. Then, the inventors have performed the BLAST search (Altschul S.F., Gish W., Miller W., Myers E.W., and Lipman D.J. (1990) J. Mol. Biol. 215: 403-410; Gish W., and States D.J. (1993) Nature Genet. 3: 266-272; http://www.ncbi.nlm.nih.gov/BLAST/) of the GenBank (http://www.ncbi.nlm.nih.gov/Web/GenBank/index.html) and SwissProt (http://www.ebi.ac.uk/ebi_docs/swissprot_db/swisshome.html) using the deduced amino acid sequence to accomplish the present invention.

The present invention relates to the polynucleotide mentioned below, a protein encoded by the polynucleotide, and their usage.

First, the present invention relates to

(1) an isolated polynucleotide selected from the group consisting of

(a) a polynucleotide comprising a coding region of the nucleotide sequence set forth in any one of the SEQ ID NOs in Table 1;

(b) a polynucleotide comprising a nucleotide sequence encoding a protein comprising the amino acid sequence set forth in any one of the SEQ ID NOs in Table 1;

(c) a polynucleotide comprising a nucleotide sequence encoding a protein comprising an amino acid sequence selected from the amino acid sequences set forth in the SEQ ID NOs in Table 1, in which one or more amino acids are substituted, deleted, inserted, and/or added, wherein said protein is functionally equivalent to the protein comprising said amino acid sequence selected from the amino acid sequences set forth in the SEQ ID NOs in Table 1;

(d) a polynucleotide that hybridizes with a polynucleotide comprising a nucleotide sequence selected from the nucleotide sequences set forth in the SEQ ID NOs in Table 1, and that comprises a nucleotide sequence encoding a protein functionally equivalent to the protein encoded by the nucleotide sequence selected from the nucleotide sequences set forth in the SEQ ID NOs in Table 1;

(e) a polynucleotide comprising a nucleotide sequence encoding a partial amino acid sequence of a protein encoded by the polynucleotide of (a) to (d);

(f) a polynucleotide comprising a nucleotide sequence with at least 70% identity to the nucleotide sequence set forth in any one of the SEQ ID NOs in Table 1.

Table 1 shows the names of the cDNA clones isolated in the examples described later, comprising the full length cDNA of the present invention, the corresponding SEQ ID NOs. of the nucleotide sequences of the cDNA clones, and the corresponding SEQ ID NOs. of the amino acid sequences deduced from the nucleotide sequences of the cDNA clones.

Table 1

Amino acid sequence	Nucleotide sequence	Clone Name
SEQ ID: 2	SEQ ID: 1	PSEC0006
SEQ ID: 4	SEQ ID: 3	PSEC0043
SEQ ID: 6	SEQ ID: 5	PSEC0058
SEQ ID: 8	SEQ ID: 7	PSEC0211

Furthermore, the present invention relates to the above polynucleotide, a protein encoded by the polynucleotide, and the use of them as described below.

(2) A substantially pure protein encoded by the polynucleotide of (1).

(3) A partial peptide of the protein of (2).

(4) An antibody against the protein of (2) or the peptide of (3).

(5) A vector comprising the polynucleotide of (1).

(6) A transformant carrying the polynucleotide of (1) or the vector of (5).

(7) A transformant expressively carrying the polynucleotide of (1) or the vector of (5).

(8) A method for producing the protein of (2) or the peptide of (3), comprising culturing the transformant of (7) and recovering the expression product.

(9) An oligonucleotide comprising the nucleotide sequence set forth in any one of the

SEQ ID NOs in Table 1 or the nucleotide sequence complementary to the complementary strand thereof, wherein said oligonucleotide comprises 15 nucleotides or more.

(10) Use of the oligonucleotide of (9) as a primer for synthesizing a polynucleotide.

(11) Use of the oligonucleotide of (9) as a probe for detecting a gene.

(12) An antisense polynucleotide against the polynucleotide of (1), or the portion thereof.

(13) A method for synthesizing a polynucleotide, the method comprising:

a) synthesizing a complementary strand using a cDNA library as a template, and using the primer of (10); and

b) recovering the synthesized product.

(14) The method of (13), wherein the cDNA library is obtainable by oligo-capping method.

(15) The method of (13), wherein the complementary strand is obtainable by PCR.

(16) A method for detecting the polynucleotide of (1), the method comprising:

a) incubating a target polynucleotide with the oligonucleotide of (9) under the conditions where hybridization occurs, and

b) detecting the hybridization of the target polynucleotide with the oligonucleotide of (9).

Any patents, patent applications, and publications cited herein are incorporated by reference.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 shows the restriction maps of vectors pME18SFL3 and pUC19FL3.

DETAILED DESCRIPTION OF THE INVENTION

Herein, "polynucleotide" is defined as a molecule in which multiple nucleotides are polymerized. There are no limitations in the number of the polymerized nucleotides. In case that the polymer contains relatively low number of nucleotides, it is also described as an "oligonucleotide". The polynucleotide or the oligonucleotide of the present invention can be a natural or chemically synthesized product. Alternatively, it can be synthesized using a template polynucleotide by an enzymatic reaction such as PCR.

All the cDNA provided by the invention are full length cDNA. Herein, a "full length cDNA" is defined as a cDNA which contains both ATG codon (the translation start site) and the stop codon. Accordingly, the untranslated regions, which are originally found in the upstream or downstream of the protein coding region in natural mRNA, may or may

not be contained.

An "isolated polynucleotide" is a polynucleotide which is not identical to any naturally occurring nucleic acid or to that of any fragment of a naturally occurring genomic nucleic acid spanning more than three separate genes. The term therefore covers, for example,

(a) a DNA which has the sequence of part of a naturally occurring genomic DNA molecule but is not flanked by both of the coding sequences that flank that part of the molecule in the genome of the organism in which it naturally occurs;

(b) a nucleic acid incorporated into a vector or into the genomic DNA of a prokaryote or eukaryote in a manner such that the resulting molecule is not identical to any naturally occurring vector or genomic DNA;

(c) a separate molecule such as a cDNA, a genomic fragment, a fragment produced by polymerase chain reaction (PCR), or a restriction fragment; and

(d) a recombinant nucleotide sequence that is part of a hybrid gene, i.e., a gene encoding a fusion protein. Specifically excluded from this definition are nucleic acids present in mixtures of different (i) DNA molecules, (ii) transfected cells, or (iii) cell clones: e.g., as these occur in a DNA library such as a cDNA or genomic DNA library.

A substantially pure human protein of the present invention comprises any one of the amino acid sequences of SEQ ID NO: 2, SEQ ID NO: 4, SEQ ID NO: 6, and SEQ ID NO: 8, as shown in Table 1. The features of these proteins and the full length cDNA clones encoding the proteins were summarized in Table 2. Among the clones, PSEC0058 has a longer 5'-end sequence than the dbEST in the GenBank.

The term "substantially pure" as used herein in reference to a given polypeptide means that the protein or polypeptide is substantially free from other biological macromolecules. The substantially pure protein or polypeptide is at least 75% (e.g., at least 80, 85, 95, or 99%) pure by dry weight. Purity can be measured by any appropriate standard method, for example, by column chromatography, polyacrylamide gel electrophoresis, or HPLC analysis.

Table 2

name of the clones	number of bases	number of amino acids	ATG No. (initiation)	ATGpr1
PSEC0006	1246bp	296aa	2	0.85
PSEC0043	1811bp	269aa	10	0.90
PSEC0058	4248bp	745aa	4	0.17
PSEC0211	1545bp	222aa	4	0.60

Since the amino acid sequence of the protein of the present invention has been determined, it is possible to analyze its biological function(s) of the clone gene by expressing it as a recombinant protein utilizing an appropriate expression system, or by using a specific antibody against it.

It is possible to analyze the function of the cloned gene, for example, by expressing the protein of the invention, injecting the protein into cells (various cell lines or primary culture cells), and analyzing the changes in cells by monitoring the changes in signals such as calcium ions, the change of the cellular growth state, or the change of the expression of a protein or mRNA whose function is known. It is also possible to analyze the function of the cloned gene by injecting into cells (various cell lines or primary culture cells) the antibody which specifically recognizes the protein of the invention, and analyzing the changes in the cells by monitoring the changes in signals such as calcium ions, the change of the cellular growth state, or the change of the expression of a protein or mRNA whose function is known. Furthermore, it is possible to predict the function of the protein of the invention by analyzing the localization of the polypeptide within the cells or within the tissues in detail by using an antibody that recognizes the protein specifically. For example, the histochemical analysis of the whole body of an embryo (in case that it is difficult to obtain a human embryo, a mouse one, for example, can be used, since the corresponding genes, for example, of mouse generally has high homology to the human genes at the amino acid level. In particular, simian genes have high homology to the human gene), cells in each differentiation level, or cultured cells can be used to predict the function of the cloned gene.

Since any protein encoded by the cDNA clone of the invention contains its full length amino acid sequence, it is possible to analyze its biological activity by expressing it as a

recombinant protein utilizing an appropriate expression system, or by using a specific antibody against it. If the protein is associated with diseases, a specific antibody obtained by using the expressed protein can be utilized to examine the relationship between the expression level or activity of the protein and a particular disease. Alternatively, it is possible to analyze the relationship between the protein and disease by using the Online Mendelian Inheritance in Man (OMIM) (http://www.ncbi.nlm.nih.gov/Omim/), the database of human genes and diseases. Proteins associated with diseases are useful in drug development since they can be utilized as a diagnostic marker, a drug that regulates the level of their expression and activity, or a target of gene therapy. Especially, the protein associated with transcription or signal transduction is extremely useful in the medicinal industry because the associations of such a protein with diseases have been reported in "Transcription factor research 1999" (Fujii, Tamura, Morohashi, Kageyama, and Satake edit. (1999) Jikken-Igaku Zoukan, Vol.17, No.3), and "Gene medical" (1999) Vol.3, No.2.

Based on the functional analysis of the above-mentioned protein, medicines can be developed as follows. If the protein is a regulatory factor of the cellular conditions such as growth and differentiation, low-molecular-weight compounds can be screened by examining the change in the cellular conditions, or the activation or repression of a particular gene in a certain cell into which the protein or the antibody of the present invention is microinjected.

The screening can be performed, for example, as follows. First, the protein of the invention is expressed and purified in a recombinant form. Then, the purified protein is microinjected into a various kind of cell lines or primary cultured cells, and the change in the cell growth and differentiation is monitored. The induction of a particular gene that is known to be involved in a certain cellular change is detected by the amounts of mRNA and protein. Alternatively, the amount of an intracellular molecule (low-molecular-weight compounds, etc.) that is changed by the function of a gene product (protein) that is known to function in a certain cellular change is used for the detection. A compound (which can be either a low-molecular-weight or high-molecular-weight compound) whose activity is to be screened can be screened by the change in cellular conditions as an index by adding the compound to the culture medium. In some cases, the screening can be achieved by merely monitoring the change of a gene product obtained in the present invention without microinjecting the gene product into a cell. The above screening enables developing a substance that activates or represses the function of a protein of the present invention, which regulates cellular conditions or functions. Such a substance is expected to be used

as medicine.

More specifically, the following method is available. First, a transformed cell line expressing the protein of the invention is obtained. Then, the transformed cell line and the untransformed original cell line are compared for the changes in the expression of a certain gene by detecting the amount of its mRNA or protein. Alternatively, the amount of an intracellular molecule (low molecular compounds, etc.) that is changed by the function of a certain gene product (protein) is used for the detection. Furthermore, the change in the expression of a certain gene is detected by introducing a fusion gene that comprises a regulatory region of the gene and a marker gene (luciferase, beta-galactosidase, etc.) into cells, expressing the protein of the invention in the cells, and estimating the activity of a marker gene product (protein). Once the screening reveals that the affected protein or gene is associated with diseases, it is possible to perform a screening for a compound or gene that is capable of regulating the expression or activity of the affected gene either directly or indirectly by utilizing the protein provided by the invention.

First, the protein of the invention is expressed and purified in a recombinant form. The affected protein or gene is also purified. Then, the binding ability of the protein of the invention to the affected protein or gene is examined. The change in the binding ability is monitored after a compound that is a candidate for an inhibitor is added to the reaction mixture. In an alternative method, a regulatory factor of the expression of the gene encoding the protein of the invention can be screened as follows. A transcription regulatory region locating in the 5'-upstream of the gene encoding the protein of the invention is obtained, and fused with a marker gene. After the fusion gene is introduced into cells, test compounds are added to the cells for screening. . The compounds obtained through such screening can be used as a drug for the diseases with which the protein of the invention is associated. Similarly, if the regulatory factor is a protein, compounds that affect the expression or activity of the protein can be used as a medicine for the diseases.

If the protein of the invention has an enzymatic activity, a screening can be performed by adding a compound to the protein of the invention and monitoring the change of the compound. In addition, the enzymatic activity can also be utilized to screen a compound that inhibits the activity of the protein. Such screening can be carried out as follows. First, the protein of the invention is expressed and purified in a recombinant form. Then, compounds are added to the purified protein, and the amounts of the compound and of the reaction products are examined. Alternatively, after a compound that is a candidate for an inhibitor is added, a compound (substrate) that reacts with the purified protein is added,

and the amounts of the substrate and of the reaction products are examined. The compounds obtained in the screening can be used as a medicine for diseases with which the protein of the invention is associated.

A specific antibody that recognizes the protein of the invention can be used to examine the relationship between the level of the expression or activity of the protein and a particular disease. It is also possible to analyze the relationship according to the methods described in "Molecular Diagnosis of Genetic Diseases" (Elles R. edit. (1996)) in the series of "Method in Molecular Biology" (Humana Press). Proteins associated with diseases are targets of screening as mentioned, and thus are very useful in developing drugs which regulate their expression and activity. Also, the proteins are useful in the medicinal industry as a diagnostic marker of the associated disease or a target of gene therapy.

Compounds isolated as mentioned above can be administered patients as it is, or after formulated into a pharmaceutical composition according to the known methods. For example, a pharmaceutically acceptable carrier or vehicle, specifically sterilized water, saline, plant oil, emulsifier, or suspending agent can be mixed with the compounds appropriately. The pharmaceutical compositions can be administered to patients by a method known to those skilled in the art, such as intraarterial, intravenous, or subcutaneous injections. The dosage may vary depending on the weight or age of a patient, or the method of administration, but those skilled in the art can choose an appropriate dosage properly. If the compound is encoded by DNA, the DNA can be cloned into a vector for gene therapy, and used for gene therapy. The dosage of the DNA and the method of its administration may vary depending on the weight or age of a patient, or the symptoms, but those skilled in the art can choose properly.

The protein of the invention can be prepared as a recombinant protein or a natural protein. The recombinant protein can be prepared, for example, by inserting the DNA encoding the protein of the invention into a vector, introducing the vector into an appropriate host cell culturing the host cell in a culture medium and purifying the protein expressed in the transformed host cell or the culture medium, as described below. The natural protein can be prepared, for example, by utilizing an affinity column to which an antibody against the protein of the invention is attached, as described below (Current Protocols in Molecular Biology (1987) Ausubel et al. edit, John Wily & Sons, Section 16.1-16.19). The antibody used for the affinity chromatography can be either a polyclonal antibody or a monoclonal antibody. Alternatively, *in vitro* translation (for

example, "On the fidelity of mRNA translation in the nuclease-treated rabbit reticulocyte lysate system." Dasso M.C., and Jackson R.J. (1989) Nucleic Acids Res. 17: 3129-3144) can be used for preparing the protein of the invention.

Proteins functionally equivalent to the proteins of the present invention can be prepared by those skilled in the art, for example, by using a method for introducing mutations into an amino acid sequence of a protein (for example, site-directed mutagenesis (Current Protocols in Molecular Biology, edit, Ausubel et al., (1987) John Wiley & Sons, Section 8.1-8.5). Besides, such proteins can be generated by spontaneous mutations. The present invention include the proteins having one or more amino acid substitutions, deletions, insertions and/or additions in the amino acid sequences of the proteins of the present invention specifically SEQ ID No. 2, 4, 6, and 8, as far as the proteins have the equivalent functions to those of the proteins identified in the EXAMPLE.

There are no limitations in the number and sites of amino acid mutations, as far as the proteins maintain their functions. The number of mutations typically falls within 10%, preferably within 5%, and more preferably within 1% of the total amino acids. From the viewpoint of maintaining the protein function, it is preferable that a substituted amino acid has a similar property to that of the original amino acid. For example, Ala, Val, Leu, Ile, Pro, Met, Phe and Trp are assumed to have similar properties to one another because they are all classified into a group of non-polar amino acids. Similaly, substitution can be performed among non-charged amino acids sach as Gly, Ser, Thr, Cys, Tyr, Asn, and Gln, acidic amino acids such as Asp and Glu, and basic amino acids such as Lys, Arg, and His.

In addition, proteins functionally equivalent to the proteins of the present invention can be isolated by using techniques of hybridization or gene amplification known to those skilled in the art. Specifically, using the hybridization technique (Current Protocols in Molecular Biology, edit, Ausubel et al., (1987) John Wiley & Sons, Section 6.3-6.4)), those skilled in the art can usually isolate a DNA highly homologous to the DNA encoding the protein identified in the below mentioned EXAMPLE based on the identified nucleotide sequence (SEQ ID No. 1, 3, 5, and 7) or a portion thereof and obtain the functionally equivalent protein from the isolated DNA. The present invention includes proteins encoded by the DNAs hybridizing with the DNAs encoding the proteins identified in the present EXAMPLE, as far as the proteins are functionally equivalent to the proteins identified in the present EXAMPLE. Organisms from which the functionally equivalent proteins are isolated include vertebrates such as human, mouse, rat, rabbit, pig and bovine, but are not limited to these animals.

Washing conditions of hybridization for the isolation of DNAs encoding the functionally equivalent proteins are usually "1 × SSC, 0.1% SDS, 37°C"; more stringent

conditions are "0.5×SSC, 0.1% SDS, 42°C"; and still more stringent conditions are "0.1× SSC, 0.1% SDS, 65°C". Alternatively, the following conditions can be given as hybridization conditions of the present invention. Namely, conditions in which the hybridization is done at "6×SSC, 40% Formamide, 25°C", and the washing at "1×SSC, 55°C" can be given. More preferable conditions are those in which the hybridization is done at "6×SSC, 40% Formamide, 37°C", and the washing at "0.2×SSC, 55°C". Even more preferable are those in which the hybridization is done at "6×SSC, 50% Formamide, 37°C", and the washing at "0.1×SSC, 62°C". The more stringent the conditions of hybridization are, the more frequently the DNAs highly homologous to the probe sequence are isolated. Therefore, it is preferable to conduct hybridization under stringent conditions. Examples of stringent conditions in the present invention are, washing conditions of "0.5×SSC, 0.1% SDS, 42°C", or alternatively, hybridization conditions of "6×SSC, 40% Formamide, 37°C", and the washing at "0.2×SSC, 55°C". However, the above-mentioned combinations of SSC, SDS and temperature conditions are indicated just as examples. Those skilled in the art can select the hybridization conditions with similar stringency to those mentioned above by properly combining the above-mentioned or other factors (for example, probe concentration, probe length and duration of hybridization reaction) that determines the stringency of hybridization.

The amino acid sequences of proteins isolated by using the hybridization techniques usually exhibit high homology to those of the proteins of the present invention. The present invention encompasses a polynucleotide comprising a nucleotide sequence that has a high identity to the nucleotide sequence of claim 1 (a). Furthermore, the present invention encompasses a peptide, or protein comprising an amino acid sequence that has a high identity to the amino acid sequence encoded by the polynucleotide of claim 1 (b). The term "high identity" indicates sequence identity of at least 40% or more; preferably 60% or more; and more preferably 70% or more. Alternatively, more preferable is identity of 90% or more, or 93% or more, or 95% or more, furthermore, 97% or more, or 99% or more. The identity can be determined by using the BLAST search algorithm.

With the gene amplification technique (PCR) (Current Protocols in Molecular Biology, edit, Ausubel et al., (1987) John Wiley & Sons, Section 6.3-6.4)) using primers designed based on the nucleotide sequence (SEQ ID No. 1, 3, 5, and 7) or a portion thereof identified in the present EXAMPLE, it is possible to isolate a DNA fragment highly homologous to the nucleotide sequence or a portion thereof and to obtain functionally equivalent protein to a particular protein identified in the EXAMPLE based on the isolated DNA fragment.

The "percent identity" of two amino acid sequences or of two nucleic acids is

determined using the algorithm of Karlin and Altschul (Proc. Natl. Acad. Sei. USA 87:2264-2268, 1990), modified as in Karlin and Altschul (Proc. Natl. Acad. Sei. USA 90:5873-5877, 1993). Such an algorithm is incorporated into the BLASTN and BLASTX programs of Altschul et al. (J. Mol. Biol.215:403-410, 1990). BLAST nucleotide searches are performed with the BLASTN program, score = 100, wordlength = 12. BLAST protein searches are performed with the BLASTX program, score = 50, wordlength = 3. When gaps exist between two sequences, Gapped BLAST is utilized as described in Altschul et al. (Nucleic Acids Res.25:3389-3402,1997). When utilizing BLAST and Gapped BLAST programs, the default parameters of the respective programs (e.g., BLASTX and BLASTN) are used. See http://www.ncbi.nlm.nih.gov.

The present invention also includes a partial peptide of the proteins of the invention. In addition, the present invention includes an antigen peptide for raising antibodies. The peptides to be specific for the protein of the invention comprise at least 7 amino acids, preferably 8 amino acids or more, and more preferably 9 amino acids or more. The peptide can be used for preparing antibodies against the protein of the invention, or competitive inhibitors of them, and also screening for a receptor that binds to the protein of the invention. The partial peptides of the invention can be produced, for example, by genetic engineering methods, known methods for synthesizing peptides, or digesting the protein of the invention with an appropriate peptidase.

The present invention also relates to a polynucleotide encoding the protein of the invention. The polynucleotide of the invention can be provided in any form as far as it encodes the protein of the invention, and thus includes cDNA, genomic DNA, and chemically synthesized DNA, etc. The polynucleotide also includes a DNA comprising any nucleotide sequence that is obtained based on the degeneracy of the genetic code, as far as it encodes the protein of the invention. The polynucleotide of the invention can be isolated by the standard methods such as hybridization using a probe polynucleotide comprising the nucleotide sequence set forth in SEQ ID NO: 1, SEQ ID NO: 3, SEQ ID NO: 5, or SEQ ID NO: 7, or the portions of them, or by PCR using primers that are synthesized based on the nucleotide sequence.

For example, 4 clones provided by the present invention, which have been isolated in the examples mentioned below, are novel and full length cDNA. All the cDNA clones provided by the invention are characterized as follows.

Namely, all the cDNA clones of the present invention comprises full length cDNA,

those obtained by the oligo-capping method, and those selected based on the features of their 5'end sequences by the score in the ATGpr (or described as ATGpr1) that predicts the full length ratio at the 5'-end. Moreover, the cDNA clones of the present invention are those in which the PSORT, which predicts the presence of signal sequences, has found no signal sequence at their 5'-ends and those which have no transmembrane region in their protein coding regions. In addition, the selected clones were found to be not identical to any human mRNA (therefore, to be novel) by the homology search for their 5'-end sequences.

The present invention also relates to a vector into which the DNA of the invention is inserted. The vector of the invention is not limited as long as it contains the inserted DNA stably. For example, if *E. coli* is used as a host, vectors such as pBluescript vector (Stratagene) are preferable as a cloning vector. To produce the protein of the invention, expression vectors are especially useful. Any expression vector can be used as far as it is capable of expressing the protein *in vitro*, in *E. coli*, in cultured cells, or *in vivo*. For example, pBEST vector (Promega) is preferable for *in vitro* expression, pET vector (Invitrogen) for *E. coli*, pME18S-FL3 vector (GenBank Accession No. AB009864) for cultured cells, and pME18S vector (Mol. Cell. Biol. (1988) 8: 466-472) for *in vivo* expression. To insert the DNA of the invention, ligation utilizing restriction sites can be performed according to the standard method (Current Protocols in Molecular Biology (1987) Ausubel et al. edit, John Wily & Sons, Section 11.4-11.11).

The present invention also relates to a transformant carrying the vector of the invention. Any cell can be used as a host into which the vector of the invention is inserted, and various kinds of host cells can be used depending on the purposes. For strong expression of the protein in eukaryotic cells, COS cells or CHO cells can be used, for example.

Introduction of the vector into host cells can be performed, for example, by calcium phosphate precipitation method, electroporation method (Current Protocols in Molecular Biology (1987) Ausubel et al. edit, John Wily & Sons, Section 9.1-9.9), lipofectamine method (GIBCO-BRL), or microinjection method, etc.

The present invention also relates to a polynucleotide which specifically hybridizes with a polynucleotide comprising the nucleotide sequence set forth in SEQ ID NO: 1, SEQ ID NO: 3, SEQ ID NO: 5, or SEQ ID NO: 7 encoding the protein of the invention, or its complementary strand, and has a length of at least 15 nucleotides. Herein, the term

"specifically hybridize" is used as to refer to hybridize with a polynucleotide of SEQ ID NO: 1, SEQ ID NO: 3, SEQ ID NO: 5, or SEQ ID NO: 7 encoding the protein of the invention, or its complementary strand, and not with polynucleotide encoding other proteins under the standard conditions for hybridization, or preferably under stringent conditions. Such polynucleotide can be used as a probe for isolation and detection of the polynucleotide of the invention, and as a primer for amplifying the polynucleotide of the present invention. As a primer, the polynucleotide usually has a length of 15 to 100 bp, and preferably has a length of 15 to 35 bp. As a probe, the polynucleotide contains the entire sequence of the polynucleotide of the invention, or at least the portion of it, and has a length of at least 15 bp.

The polynucleotide of the present invention can be used for examination and diagnosis of the abnormality of the protein of the invention. For example, it is possible to examine the abnormal expression of the gene encoding the protein using the polynucleotide of the invention as a probe for Northern hybridization or as a primer for RT-PCR. Also, the polynucleotide of the invention can be used as a primer for polymerase chain reaction (PCR) such as the genomic DNA-PCR, and RT-PCR to amplify the polynucleotide encoding the protein of the invention, or the regulatory region of the expression, with which it is possible to examine and diagnose the abnormality of the sequence by RFLP analysis, SSCP, and direct sequencing, etc.

Furthermore, the "polynucleotide which specifically hybridizes with a polynucleotide comprising the nucleotide sequence set forth in SEQ ID NO: 1, SEQ ID NO: 3, SEQ ID NO: 5, or SEQ ID NO: 7 encoding the protein of the invention, or its complementary strand, and has a length of at least 15 nucleotides" includes an antisense polynucleotide for blocking the expression of the protein of the invention. To exert the antisense effect, the antisense polynucleotide has a length of at least 15 bp or more, preferably 100 bp, and more preferably 500 bp or more, and has a length of usually 3000 bp or less and preferably 2000 bp or less. The antisense polynucleotide can be used in the gene therapy of the diseases which are caused by the abnormality of the protein of the invention (abnormal function or abnormal expression). Said antisense polynucleotide can be prepared, for example, by the phosphorothioate method ("Physicochemical properties of phosphorothioate oligodeoxynucleotides." Stein (1988) Nucleic Acids Res. 16: 3209-3221) based on the nucleotide sequence of the polynucleotide encoding the protein (for example, the polynucleotide set forth in SEQ ID NO: 1, SEQ ID NO: 3, SEQ ID NO: 5, or SEQ ID NO: 7).

The polynucleotide or antisense polynucleotide of the present invention can be used in gene therapy, for example, by administrating it into a patient by the *in vivo* or *ex vivo* method with virus vectors such as retrovirus vectors, adenovirus vectors, and adeno-associated virus vectors, or non-virus vectors such as liposome.

The present invention also relates to antibodies that bind to the protein of the invention. There are no limitations in the form of the antibodies of the invention. They include polyclonal antibodies, monoclonal antibodies, or their portions that can bind to the protein of the invention. They also include antibodies of all classes. Furthermore, special antibodies such as humanized antibodies are also included.

The polyclonal antibody of the invention can be obtained according to the standard method by synthesizing an oligopeptide corresponding to the amino acid sequence and immunizing rabbits with the peptide (Current Protocols in Molecular Biology (1987) Ausubel et al. edit, John Wily & Sons, Section 11.12-11.13). The monoclonal antibody of the invention can be obtained according to the standard method by purifying the protein expressed in *E. coli*, immunizing mice with the protein, and producing a hybridoma cell by fusing the spleen cells and myeloma cells (Current Protocols in Molecular Biology (1987) Ausubel et al. edit, John Wily & Sons, Section 11.4-11.11).

The antibody binding to the protein of the present invention can be used for purification of the protein of the invention, and also for detection and/or diagnosis of the abnormalities of the expression and structure of the protein. Specifically, proteins can be extracted, for example, from tissues, blood, or cells, and the protein of the invention is detected by Western blotting, immunoprecipitation, or ELISA, etc. for the above purpose.

Furthermore, the antibody binding to the protein of the present invention can be utilized for treating the diseases that associates with the protein of the invention. If the antibodies are used for treating patients, human antibodies or humanized antibodies are preferable in terms of their low antigenicity. The human antibodies can be prepared by immunizing a mouse whose immune system is replaced with that of human ("Functional transplant of megabase human immunoglobulin loci recapitulates human antibody response in mice" Mendez M.J. et al. (1997) Nat. Genet. 15: 146-156). The humanized antibodies can be prepared by recombination of the hypervariable region of a monoclonal antibody (Methods in Enzymology (1991) 203: 99-121).

The invention is illustrated more specifically with reference to the following examples, but is not to be construed as being limited thereto.

The present invention has provided 4 novel proteins and full length cDNA clones encoding the proteins. It is of great significance that the present invention has provided novel full length cDNA of human, since only few full length cDNA of human has been isolated. Since the full length cDNA clones of the present invention were derived from human, these can be associated with human diseases. The genes and proteins associated with diseases are useful as diagnostic markers. In addition, they are useful in medical development as probes for searching a compound that regulates their expression and activities, or as targets of gene therapy.

EXAMPLE 1

Construction of a cDNA library by the oligo-capping method

The NT-2 neuron progenitor cells (Stratagene), a teratocarcinoma cell line from human embryo testis, which can differentiate into neurons by treatment with retinoic acid were used. The NT-2 cells were cultured according to the manufacturer's instructions as follows.

(1) NT-2 cells were cultured without induction by retinoic acid treatment (NT2RM1).

(2) After cultured, NT-2 cells were induced by adding retinoic acid, and then were cultured for 48 hours (NT2RP1).

(3) After cultured, NT-2 cells were induced by adding retinoic acid, and then were cultured for 2 weeks (NT2RP2).

The cells were harvested separately, from which mRNA was extracted by the method described in the literature (Molecular Cloning 2nd edition. Sambrook J., Fritsch, E.F., and Maniatis T. (1989) Cold Spring Harbor Laboratory Press). Furthermore, poly(A)$^+$RNA was purified from the mRNA using oligo-dT cellulose.

Similarly, human embryo-derived tissues that were enriched with brain (HEMBA1) were used to extract mRNA by the method described in the literature (Molecular Cloning 2nd edition. Sambrook J., Fritsch, E.F., and Maniatis T. (1989) Cold Spring Harbor Laboratory Press). Furthermore, poly(A)$^+$RNA was purified from the mRNA using oligo-dT cellulose.

Each poly(A)$^+$RNA was used to construct a cDNA library by the oligo-capping method (Maruyama M. and Sugano S. (1994) Gene 138: 171-174). Using the Oligo-cap linker (SEQ ID NO: 9) and the Oligo-dT primer (SEQ ID NO: 10), bacterial alkaline phosphatase (BAP) treatment, tobacco acid phosphatase (TAP) treatment, RNA ligation, the first strand

cDNA synthesis, and removal of RNA were performed as described in the reference (Suzuki and Kanno (1996) Protein Nucleic acid and Enzyme. 41: 197-201; Suzuki Y. et al. (1997) Gene 200: 149-156). Next, 5'- and 3'-PCR primers (SEQ ID NO: 11, and 12, respectively) were used for performing PCR to convert the cDNA into double stranded cDNA, which was then digested with SfiI. Then, the DraIII-cleaved pUC19FL3 vector (Figure 1; for NT2RM1, and NT2RP1), or the DraIII-cleaved pME18SFL3 (Figure 1) (GenBank AB009864, expression vector; for NT2RP2, HEMBA1) was used for cloning the cDNA in a unidirectional manner, and cDNA libraries were obtained. The clones having an insert cDNA with a length of 1 kb length or less were discarded from the cDNA libraries. Then, the nucleotide sequence of the 5'- and 3'- ends of the cDNA clones was analyzed with a DNA sequencer (ABI PRISM 377, PE Biosystems) after sequencing reactions were performed with the DNA sequencing reagents (Dye Terminator Cycle Sequencing FS Ready Reaction Kit, dRhodamine Terminator Cycle Sequencing FS Ready Reaction Kit, or BigDye Terminator Cycle Sequencing FS Ready Reaction Kit, from by PE Biosystems) according to the instructions.

The full length-enriched cDNA libraries of NT2RP2 and HEMBA1 were constructed using eukaryotic expression vector pME18SFL3. The vector contains SRα promoter and SV40 small t intron in the upstream of the cloning site, and SV40 polyA added signal sequence site in the downstream. As the cloning site of pME18SFL3 has asymmetrical DraIII sites, and the ends of cDNA fragments contain SfiI sites complementary to the DraIII sites, the cloned cDNA fragments can be inserted into the downstream of the SRα promoter unidirectionally. Therefore, clones containing full length cDNA can be expressed transiently by introducing the obtained plasmid directly into COS cells. Thus, the clones can be analyzed very easily in terms of the proteins that are the gene products of the clones, or in terms of the biological activities of the proteins.

The fullness ratio at the 5'-end sequences of the cDNA clones in the libraries constructed by the oligo-capping method was determined as follows. Of all the clones whose 5'-end sequences were found in those of known human mRNA in the public database, a clone was judged to be "full length", if it had a longer 5'-end sequence than that of the known human mRNA, or, even though the 5'-end sequence was shorter, if it contained the translation initiation codon. A clone which did not contain the translation initiation codon was judged to be "not-full length". The fullness ratio ((the number of full length clones)/(the number of full length and not-full length clones)) at the 5'-end of the cDNA clones from each library was determined by comparing with the known human mRNA (NT2RM1:

69%; NT2RP1: 75%; NT2RP2: 62%; HEMBA1: 53%). The result indicates that the fullness ratio at the 5'-end sequence was extremely high.

The relationship between the cDNA libraries and the clones is shown below.

NT2RM1 : PSEC0006

NT2RP1 : PSEC0043

NT2RP2 : PSEC0058

HEMBA1 : PSEC0211

EXAMPLE 2

Estimation of the fullness ratio at the 5'-end of the cDNA by the ATGpr and the ESTiMateFL

The ATGpr, developed by Salamov A.A., Nishikawa T., and Swindells M.B. in the Helix Research Institute, is a program for prediction of the translation start codon based on the characteristics of the sequences in the vicinity of the ATG codon. The results are shown with expectations (also described as ATGpr1 below) that an ATG is a true start codon (0.05-0.94). When the program was applied to the 5'-end sequences of the clones from the cDNA library that was obtained by the oligo-capping method and that had 65% fullness ratio, the sensitivity and specificity of evaluation of a full length clone (clone containing the N-terminal end of ORF) were improved to 82-83% by selecting only clones having the ATGpr1 score 0.6 or higher. Furthermore, the 17,365 clones in which the 5'-end sequence is identical to a known human mRNA and which were cloned from the human cDNA libraries constructed by the oligo-capping method, were estimated by the program. Briefly, the maximal ATGpr1 score of the clones was determined, and then their 5'-end sequence was compared with the known human mRNA to estimate whether the clone is full length or not. The result was summarized in Table 3. It is indicated that the method for the selection through the combination of the ATGpr and the clones isolated from the human cDNA library that was constructed by the oligo-capping method was very efficient.

Table 3

maximal ATGpr1 Score	number of full length and not-full length clones	number of full length clones	fullness ratio
>=0.70	10,226	8,428	82.4%
>=0.50	12,171	9,422	77.4%
>=0.30	14,102	10,054	71.3%
>=0.17	15,647	10,385	66.4%
>=0.05	17,365	10,608	61.1%

** number of full length clones, the number of the clones which contain the N-terminus of the ORF; the number of not-full length clones, number of the clones which does not contain the N-terminus of the ORF; fullness ratio, the resulting number of (the number of full length clones)/(the number of full length and not-full length clones)

The ESTiMateFL, developed by Nishikawa and Ota in the Helix Research Institute, is a method for the selection of a clone with high fullness ratio by comparing with the 5'-end or 3'-end sequences of ESTs in the public database.

By the method, a cDNA clone is judged presumably not to be full length if there exist any ESTs which have longer 5'-end or 3'-end sequences than the clone. The method is systematized for high throughput analysis. A clone is judged to be full length if the clone has a longer 5'-end sequence than ESTs in the public database. Even if a clone has a shorter 5'-end, the clone is judged to be full length if the difference in length is within 50 bases, and otherwise judged not to be full length, for convenience. The accuracy of the prediction by comparing cDNA clones with ESTs is improved with increasing number of ESTs to be compared. However, when only a limited number of ESTs are available, the reliability becomes low. Thus, the method is effective in excluding clones with high probability of being not-full length, from the cDNA clones that is synthesized by the oligo-capping method and that have the 5'-end sequences with about 60 % fullness ratio. In particular, the ESTiMateFL is efficiently used to estimate the fullness ratio at the 3'-end sequence of cDNA of a human unknown mRNA which has a significant number of ESTs in the public database.

The results were summarized in Tables 4 and 5. It was confirmed that, in estimating the fullness ratio at the 5'-end sequence of the clones of the human cDNA library that was constructed by the oligo-capping method, the fullness ratio was improved even for the clones having low score in the ATGpr by combining the ATGpr and ESTiMateFL. The result was applied to the estimation of the fullness ratio at the 5'-end sequence of the clones whose complete cDNA sequence were determined. The number of full length clones, the number of not-full length clones, and the fullness ratio indicate the number of the clones which contain the N-terminus of the ORF, the number of the clones which does not contain the N-terminus of the ORF, and the resulting number of (the number of full length clones)/(the number of full length and not-full length clones), respectively.

Table 4

The fullness ratio at the 5'-end sequence of the cDNA clones that were judged to be full length by comparing the ORF of the known human mRNA and that were obtained by the oligo-capping method, wherein the ratio was evaluated by comparing the cDNA clones with ESTs.

maximal ATGpr1 Score	number of full length clones	number of not-full length clones	fullness ratio
>=0.30	8,646	907	90.5%
>=0.17	10,158	1,150	89.8%
>=0.05	15,351	2,728	84.9%

Table 5

The fullness ratio at the 5'-end sequence of the cDNA clones that were judged to be not-full length by comparing the ORF of the known human mRNA and that were obtained by the oligo-capping method, wherein the ratio was evaluated by comparing the cDNA clones with ESTs.

maximal ATGpr1 Score	number of full length clones	number of not-full length clones	fullness ratio
>=0.30	1,271	2,156	37.1%
>=0.17	1,678	2,907	36.6%
>=0.05	2,512	4,529	35.7%

EXAMPLE 3

Selection of clones with high fullness ratio and the complete DNA sequencing

Among the clones of the present invention, PSEC0006-PSEC0058 were selected by the presence of an ORF (Open reading frame: a region translated into amino acids) in the 5'-end sequence. However, the clones were not selected by the ATGpr score of the data of the 5'-end sequence (one pass sequencing). In addition, PSEC0211 was selected as those having the maximal ATGpr1 score 0.7 or higher, and containing an ORF in the 5'-end sequence.

For the selected 4 clones, the nucleotide sequences of the full length cDNA and the deduced amino acid sequences were determined. The nucleotide sequences were finally determined by overlapping completely the partial nucleotide sequences determined by the three methods mentioned below. The amino acid sequences were deduced from the determined cDNA sequences. The results were shown in SEQUENCE LISTING.

(1) Long-read sequencing from both ends of the cDNA inserts using a Licor DNA sequencer (After sequence reactions were performed according to the manual for the Licor sequencer (Aroka), DNA sequence was determined by the sequencer.)

(2) Nested sequencing by the Primer Island method which utilizes the *in vitro* transfer of AT2 transposon (Devine S.E., and Boeke J.D. (1994) Nucleic Acids Res. 22: 3765-3772) (After clones were obtained using a kit from PE Biosystems, sequence reactions were performed using the DNA sequencing reagents from the company, according to the manufacturer's instructions, and DNA sequence was determined using an ABI PRISM 377 sequencer.)

(3) Primer walking by the dideoxy terminator method using custom synthesized DNA primers (After sequence reactions were performed using the DNA sequencing reagents from PE Biosystems and custom synthesized DNA primers according to the manufacturer's instructions, DNA sequence was determined using an ABI PRISM 377 sequencer).

The sequences were subjected to the analysis by the ATGpr and the BLAST search of the GenBank and SwissProt databases. As a result, no known amino acid sequence that has high homology to any of the four clones isolated above was found. In addition, no characteristic motif was found in the clones.

Thus, the four clones (PSEC0006, PSEC0043, PSEC0058, and PSEC0211) are defined as "the clones that are predicted to be full length cDNA clones by the ATGpr, etc. among a

human cDNA library that was constructed by the oligo-capping method and that has high full length ratio." Among the clones, PSEC0058, which has low score in the ATGpr1 (ATGpr1 0.17), is a clone obtained by the complete DNA sequencing of the clones that were selected as those having a long ORF based on the data of 5'-end sequence of the cDNA (one pass sequencing). Namely, PSEC0058 is not a clone selected by the ATGpr. The comparison between PSEC0058 and corresponding ESTs revealed that PSEC0058 is longer than any of the ESTs.